A Semantic Framework for Web-based Accommodation Information Integration

Kai Yang

University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2012

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

_____

Abstract

With the tremendous growth of the Web, a broad spectrum of accommodation information is to be found on the Internet. In order to adequately support information users in collecting and sharing information online, it is important to create an effective information integration solution, and to provide integrated access to the vast numbers of online information sources. In addition to the problem of distributed information sources, information users also need to cope with the heterogeneous nature of the online information sources, where individual information sources are stored and presented following their own structures and formats. In this thesis, we explore some of the challenges in the field of information integration, and propose solutions to some of the arising challenges. We focus on the utilization of ontology for integrating heterogeneous, structured and semi-structured information sources, where instance level data are stored and presented according to meta-data level schemas. In particular, we looked at XML-based data that is stored according to XML schemas.

In a first step towards a large-scale information integration solution, we propose a semantic integration framework. The proposed framework solves the problem of information integration on three levels: the data level, process level and architecture level. On the data level,

we leverage the benefit of ontology, and use ontology as a mediator for enabling semantic interoperability among heterogeneous data sources. On the process level, we alter the process of information integration, and propose a three step integration process named as the publish-combine-use mechanism. The primary goal is to distribute the efforts of collecting and integrating information sources to various types of end users. In the proposed approach, information providers have more control over their own data sources, as data sources are able to join and leave the information sharing network according to their own preferences. On the architecture level, we combine the flexibility offered by the emerging distributed P2P approach with the query processing capability provided by the centralized approach. The joint architecture is similar to the structure of the online accommodation industry.

This thesis also demonstrates the practical applicability of the proposed semantic integration framework by implementing a prototype system. The prototype system named the "accommodation hub" is specifically developed for integrating online accommodation information in the large, distributed, heterogeneous online environment. The proposed semantic integration solution and the implemented prototype system are evaluated to provide a measure of the system performance and usage. Results show that the proposed solution delivers better performance with respect to some of the evaluation criteria than some related approaches in information integration.

Acknowledgements

This thesis is fully supported by the ARC Linkage Project, a joint effort between the University of Technology, Sydney and the Lido Group. Here, I would like to acknowledge that this research would not have been possible without the support from both the university and the industry partner.

First of all, I want to thank my supervisor, Prof. Robert Steele, who has brought me to this exciting research project, and guided me throughout my candidature. I would like to express my gratitude to him for providing me with the inspirational ideas and comments for this research.

I also wish to thank all my colleagues from the university of technology, Sydney for the invaluable inputs to this research and the enjoyable teamwork. In particular, I would like to express my gratitude to Amanda Lo for the productive collaboration and the delightful company throughout this research. Thanks to John Murphy and Rita Shen for the valuable discussions had within this research. I am also grateful to our industry partner, more specifically Matthew Tyler and John Taranto for all their advices and feedbacks on the prototype system.

Finally, my gratitude to my family and friends who made this work possible with their support and encouragement.

Contents

List of Figures

List of Tables